Medical Image Retrieval and Automatic Annotation: OHSU at ImageCLEF 2007

Jayashree Kalpathy-Cramer and William Hersh

Department of Medical Informatics & Clinical Epidemiology
Oregon Health and Science University, Portland, OR, USA
{kalpathy,hersh}@ohsu.edu

Abstract. Oregon Health & Science University participated in the medical retrieval and medical annotation tasks of ImageCLEF 2007. In the medical retrieval task, we created a web-based retrieval system built on a full-text index of both image and case annotations. The text-based search engine was implemented in Ruby using Ferret, a port of Lucene and a custom query parser. In addition to this textual index of annotations, supervised machine learning techniques using visual features were used to classify the images based on image acquisition modality. All images were annotated with the purported modality. Purely textual runs as well as mixed runs using the purported modality were submitted, with the latter performing among the best of all participating research groups. In the automatic annotation task, we used the 'gist' technique to create the feature vectors. Using statistics derived from a set of multi-scale oriented filters, we created a 512-dimensional vector. PCA was then used to create a 100-dimensional vector. This feature vector was fed into a two layer neural network. Our error rate on the 1000 test images was 67.8 using the hierarchical error calculations.

1 Medical Image Retrieval

Advances in digital imaging technologies and the increasing prevalence of Picture Archival and Communication Systems (PACS) have led to a substantial growth in the number of digital images stored in hospitals and medical systems in recent years. In addition, on-line atlases of images have been created for many medical domains including dermatology, radiology and gastroenterology. Medical images can form an essential component of a patient's health record. Medical image retrieval systems can be important with aiding in diagnosis and treatment. They can also be highly effective in health care education, for students, instructors and patients.

1.1 Introduction

Image retrieval systems do not currently perform as well as their text counterparts [1]. Medical and other image retrieval systems have historically relied on annotations or captions associated with the images for indexing the retrieval system. The last few decades have seen numerous advancements in the area of content-based image retrieval (CBIR) [2,3]. Although CBIR systems have demonstrated success in fairly

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 623–630, 2008.

constrained medical domains including pathology, dermatology, chest radiology, and mammography, they have demonstrated poor performance when applied to databases with a wide spectrum of imaging modalities, anatomies and pathologies [1,4,5,6].

Retrieval performance has shown demonstrable improvement by fusing the results of textual and visual techniques. This has especially been shown to improve early precision [7,8]. The medical image retrieval task within ImageCLEF (ImageCLEFmed) 2007 campaign is a TREC-style [9] and provides a forum and set of test collections for the medical image retrieval community to use to benchmark their algorithms on a set of queries. The ImageCLEF campaign has, since 2003, been a part of the Cross Language Evaluation Forum (CLEF) [9,10,11] which is derived from the Text Retrieval Conference (TREC, trec.nist.gov).

1.2 System Description of Our Adaptive Medical Image Retrieval System

The ImageCLEF collection consists of about 66,000 medical images and annotations associated with them. We have created a flexible database schema that allows us to easily incorporate new collections while facilitating retrieval using both text and visual techniques. The text annotations in the collection are currently indexed and we continue to add indexable fields for incorporating visual information.

Database and Web Application. We used the Ruby programming language, with the open source Ruby On Rails web application framework[1, 2]. A PostgreSQL relational database was used to store the images and annotations.

The database has images from the four different collections that were part of the ImageCLEFmed 2006 image retrieval challenge as well as two new collections for 2007. The approximately 66,000 images in these collections reside in cases, with annotations in English, German and/or French. The collections themselves are substantially heterogeneous in their architectures. Some collections have only one image per case while others have many images per case. Annotation fields are also quite different among the collections. Some collections have case-based annotations while others have image-based annotations. This difference is especially significant for text based retrieval as images of different modalities or anatomies or pathologies could be linked to the same case annotation. In this situation, even though only one image from a case containing many images might be relevant to a query (based on the annotation), all images for the case would be retrieved in a purely text based system, reducing the precision of the search.

We used the relational database to maintain the mappings between the collections, the cases in the collections, the cased-based annotations, the images associated with a collection, and the image based annotations.

Image Processing and Analysis. The image itself has important visual characteristics such as color and texture that can help in the retrieval process. Images that may have had information about the imaging modality or anatomy or view associated with them as part of the DICOM header can lose that information when the image is compressed

[1] http://www.ruby-lang.org

[2] http://www.rubyonrails.org

to become a part of a teaching or on-line collection, as the image format used by these collections is usually compressed JPEG.

We created additional tables in the database to store image information that was created using a variety of image processing techniques in MATLAB[3]. For instance, the images in the collection typically do not contain explicit details about the imaging modality. In previous work [8], we have described our modality classifier that can identify the imaging modality for medical images with a high level of confidence (>95% accuracy on the database used for the validation). Grey scale images are classified into a set of modalities including x-rays, CT, MRI, ultrasound and nuclear medicine. Color image classes include gross pathology, microscopy, and endoscopy.

Each image was annotated in the database with the purported image modality and a confidence value. This can be extremely useful for queries where the user has specified a desired image modality. An example query from ImageCLEF 2006 was "*Show me microscopic images of tissue from the cerebellum.*"

The precision of the result of such a query can be improved significantly by restricting the images returned to those of the modality desired [8]. This is especially useful in eliminating images of the incorrect modality that may be part of a case containing a relevant image from the returned list of images. However, this increase in precision may result in a loss in recall if the classification algorithm incorrectly classifies the image modality.

We continue to experiment with a variety of image clustering and classification algorithms and adding the numerical data and labels to the database. Clustering images that look visually similar can be again used to improve the precision of the image retrieval process and speed up the system searching of images in the same cluster as the query image (if available).

Query Parser and Search Engine . The system presents search options to the user including Boolean OR, AND and exact match. There are also options to perform fuzzy searches and custom query parsing. The cornerstone of our system is the query parser, written in Ruby. Ferret, a Ruby port of the popular Lucene system, was used in our system as the underlying search engine[4].

Queries were first analyzed using MedPost, a Parts-of-Speech (POS) Tagger created using the Medline corpus, and distributed by the National Library of Medicine[5] [14].

A simple Bayesian classifier[6] was trained to discern the desired image modality from the query, if available. The classifier performed extremely well within the constrained vocabulary of imaging modalities. Stop words were then removed from the query. These include Standard English stop words as well as a small set of stop words determined by analyzing queries from the last three years, including 'finding', 'showing', 'images', 'including' and 'containing'.

The system is also linked to the UMLS Metathesaurus. The user can choose to perform automatic query expansion using synonyms from the Metathesaurus.

[3] http://www.mathworks.com

[4] http://ferret.davebalmain.com

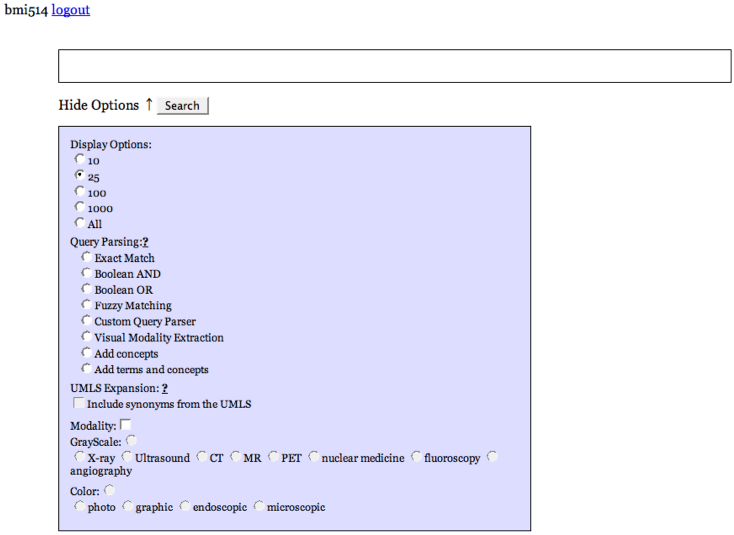
[5] ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedPost/medpost

[6] http://classifier.rubyforge.org

A sample query "Show me CT images with a brain infarction" is automatically parsed and the following information is extracted from it: CT-> imaging modality, brain -> anatomic location, infarction -> finding. This information can be used to combine the results of the textual and visual systems more effectively.

1.3 Runs Submitted

We submitted a total of 10 runs. These runs included textual and mixed, automatic and manual options. The text runs had an "as-is" run, where the topics were submitted directly to the search system, a run where term expansion using the UMLS system was used, a text run where both our custom parser and query expansion was used and a manual run. We also submitted runs using different weighted combinations of the FIRE baseline (published by the organizers) with our baseline textual runs.



bmi514 logout

Hide Options ↑ Search

Display Options:
○ 10
● 25
○ 100
○ 1000
○ All

Query Parsing:?
○ Exact Match
○ Boolean AND
○ Boolean OR
○ Fuzzy Matching
○ Custom Query Parser
○ Visual Modality Extraction
○ Add concepts
○ Add terms and concepts

UMLS Expansion: ?
□ Include synonyms from the UMLS

Modality: □
GrayScale: ○
○ X-ray ○ Ultrasound ○ CT ○ MR ○ PET ○ nuclear medicine ○ fluoroscopy ○ angiography
Color: ○
○ photo ○ graphic ○ endoscopic ○ microscopic

Fig. 1. Screen display of our system displaying user options

1.4 Results and Discussion

The complete performance of our runs can be found among the official ImageCLEFmed results. However, we note that there was a discrepancy between the order in the output of our image retrieval system and that which is required for trec_eval. In calculating the mean average precision (MAP), trec_eval only considers the "similarity score" column. Both the rank column and the order of the documents in the submission are ignored. Ties are broken lexicographically. Many participants, including OHSU, had created an ordered list of images, where the order in which the documents (images) appeared was considered the ranking of the documents. However, the score was either increasing from the top of the list to the bottom of the list or a number that was not unique or indicative of the desired ranking. This poor formatting of the submissions led to surprisingly poor performance of some combined runs as well as runs of certain participants. This was discovered during the post workshop analysis of the results. The official runs, including the OHSU runs were reformatted where the score was set equal to 1/row order. This ensured that the score was in decreasing order from the top of the ordered list to the bottom of the ordered list, as required by trec_eval. Table 1 presents the official mean average precision (MAP) as well as the results of trec_eval on the reformatted runs for the most significant runs submitted by OHSU.

Table 1. Performance of significant OHSU runs

Run	Type	Official MAP	Reformatted MAP	Comments
ohsu_m2_rev1_c.txt	AM	0.341	0.408	mixed run using modality, starting from OHSU_txt_exp2
OHSU-oshu_man2	MT	0.346	0.360	manual run, using terms from umls expansion
ohsu_text_e4_out_rev1_c.txt	AT	0.332	0.347	query expansion and query parsing
OHSU-OHSU_txt_exp2	AT	0.319	0.334	query expansion using UMLS
OHSU-oshu_as_is_1000	AT	0.275	0.281	standard input with additional stop words

Our baseline textual run had a better than average performance, with a MAP of 0.28. The use of query expansion with UMLS synonyms as well as query parsing further improved the MAP. However, the most notable improvement was with the use of our modality classifier. By incorporating visual information, the MAP increases to 0.408, which is significantly better than any other official run submitted.

1.5 Conclusions and Future Work

Our image retrieval system built using open-source tools is a flexible platform for evaluating various tools and techniques in image processing as well as natural language processing for medical image retrieval. The use of visual information to automatically extract the imaging modality is a promising approach for the ImageCLEFmed campaign. The use of UMLS term expansion, query parsing and modality detection all add value over the basic Ferret (Lucene) search engine. We will continue to improve our image retrieval system by adding more image tags using automatic visual feature extraction. Our next goal is to annotate the images with the their anatomical location and view attributes.

2 Automatic Image Annotation Task

The goal of this task was to correctly classify 1000 radiographic medical images using the hierarchical IRMA code. This code classifies the image along the modality, body orientation, body region, and biological system axes. There were 116 unique classes. The task organizers provided a set of 9,000 *training* images and 1000 *development* images. The goal of the task was to classify the images to the most precise level possible, with a greater penalty applied for incorrect classification than for a less specific classification in the hierarchy.

2.1 Introduction

A supervised machine learning approach using global gist features and neural network architecture was employed for the task of automatic annotation of medical images with the IRMA code.

2.2 System Description

The automatic image annotation was based on a neural network classifier using Gist features [14]. The classifiers were created in MATLAB using the Netlab toolbox [15]. All images were convolved with a set of 32 multiscale-oriented Gabor filters. We created a 512-dimensional vector using statistics from these filters. Principal component analysis was then used to reduce the dimensionality of the vector to 100. A multilayer perceptron with one hidden layer containing 250-500 nodes was used to create and train a multi-class classifier. The training data set of 10,000 images was used to optimize performance of the development set of 1000 images. The final configuration of the classifier used 300 hidden nodes.

A confusion matrix was used to identify the most common mode of misclassification. We noted that classes 1123-110-500-000 (108) and 1123-127-500-000 (111) were frequently interchanged by our classifier. This error arises from the similarity of the Anterior-Posterior (AP) and the PA views of chest x-rays. To handle this special case, we created a second layer of classification built around a support vector machine (SVM) using scale-invariant feature transform (SIFT) features [16] as inputs. This new binary classify was used to determine the final class assignments for images in classes 108 and 111.

2.3 Runs Submitted

OHSU submitted two runs for the automatic annotation task. The first run used gist feature vectors to train the multi-layer perceptron. A neural network was used to create a multi-class classifier consisting of 116 classes. These were the original classes from 2006 and did not use the hierarchical nature of the IRMA code. These classes were then converted to the IRMA code, as required for the submission in 2007. The second run used a hierarchical classifier architecture, with the first layer as described above and the second classifier using SIFT features and an SVM.

2.4 Results and Analysis

The relationship between semantic and visual hierarchy remains an open area of research. Based on our experiments using this collection of images used for automatic annotation, the use of hierarchy of the semantic classes did not improve our automatic annotations as visual hierarchy did not correspond to semantic hierarchy.

The error count for both our runs were quite similar at 67.8 and 67.97 for 1000 images, compared to the best count of 26.84 and worst count of 505.61. There was only a very slight improvement in using the two-layer classifier. There were 227 errors using the 2006 classes, which corresponds to an classification accuracy of 77.3%. However, of these 227 errors, only 15 were wrong along all 4 axes. 76 were misclassified along two axes (primarily view and anatomy) while 12 were misclassified along 3 axes. 77 of our single misclassifications were along the view axis. A significant portion of these occurred where class 111 was misclassified as 108, an error due to confusion between posterior-anterior and anterior-posterior views of the chest.

2.5 Future Work

We would like to further investigate the mapping between the semantic and visual hierarchy of images in the IRMA collection. We primarily used a flat classifier in this work, with a constant cost for all classes and misclassifications. However, it might be possible to improve the performance using the IRMA hierarchy by the use of a cost function that depends on the hierarchy of the IRMA classes.

Acknowledgments

We acknowledge the support of NLM Training Grant 1T15 LM009461 and NSF Grant ITR-0325160. We would also like to thank Steven Bedrick for his help in creating the web-based image retrieval system.

References

1. Hersh, W., Muller, H., et al.: Advancing biomedical image retrieval: development and analysis of a test collection. J. Am. Med. Inform. Assoc. 13(5), 488–496 (2006)
2. Smeulders, A.W.M., Worring, M., et al.: Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)
3. Tagare, H.D., Jaffe, C., et al.: Medical Image Databases: A Content-Based Retrieval Approach. J. Am. Med. Inform. Assoc. 4(3), 184–198 (1997)
4. Aisen, A.M., Broderick, L.S., et al.: Automated storage and retrieval of thin-section CT images to assist diagnosis: System description and preliminary assessment. Radiology 228, 265–270 (2003)
5. Schmid-Saugeon, P., Guillod, J., et al.: Towards a computer-aided diagnosis system for pigmented skin lesions. Computerized Medical Imaging and Graphics 27, 65–78 (2003)
6. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medicine – clinical benefits and future directions. Int. J. Med. Inform. 73, 1–23 (2004)
7. Hersh, W., Kalpathy-Cramer, J., et al.: Medical image retrieval and automated annotation: OHSU at ImageCLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 660–669. Springer, Heidelberg (2007)

8. Kalpathy-Cramer, J., Hersh, W.: Automatic Image Modality Based Classification and Annotation to Improve Medical Image Retrieval. In: MedInfo 2007, Brisbane, Australia, pp. 1334–1338 (2007)
9. Braschler, M., Peters, C.: Cross-language evaluation forum: objectives, results, achievements. Inform Retriev (7), 7–31 (2004)
10. Müller, H., Deselaers, T., Lehmann, T., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2006 medical retrieval annotation tasks. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 595–608. Springer, Heidelberg (2007)
11. Müller, H., Clough, P., et al.: Evaluation Axes for Medical Image Retrieval Systems - The ImageCLEF Experience. In: ACM Int. Conf. on Multimedia, Singapore (November 2005)
12. Florea, F., Müller, H., Rogozan, A., Geissbühler, A., Darmoni, S.: Medical image categorization with MedIC and MedGIFT. In: Medical Informatics Europe (MIE 2006) (2006)
13. Smith, L., Rindflesch, T., Wilbur, W.: MedPost: a part-of-speech tagger for biomedical text Bioinformatics. 20(14) (2004)
14. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Computer Vision 42(3), 145–175 (2001)
15. Nabney, I.T.: Netlab: Algorithms for Pattern Recognition. Springer, London (2004)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. of Computer Vision 60(2), 91–110 (2004)